Claims

What is claimed is:

1. A computer-based method of mining one or more patterns in an input data set of items, the method comprising the steps of:

identifying one or more sets of items in the input data set as one or more patterns based on respective comparisons of conditional probability values associated with each of the one or more sets of items to a predetermined threshold value; and

outputting the one or more identified patterns based on results of the comparisons.

2. The method of claim 1, wherein the identifying step further comprises identifying a set of items in the input data set, which includes at least two subsets of at least one item, as a pattern when the set of items has a conditional probability value computed therefor that is not less than a predetermined threshold value, wherein the conditional probability value is indicative of a probability that both of the at least two subsets of at least one item will occur given that one of the at least two subsets of at least one item has occurred.

3. The method of claim 2, wherein the predetermined threshold value is between zero and one.

4. The method of claim 2, wherein the conditional probability value is estimated by the number of occurrences in the input data set of the two subsets divided by the number of occurrences in the data set of one of the two subsets.

5. The method of claim 1, wherein the identifying step further comprises identifying a set of items in the input data set as a pattern when the set of items has a conditional probability value computed for the set of items minus a particular item of the

set, given the particular item of the set, that is not less than a predetermined threshold value.

6. The method of claim 5, wherein the predetermined threshold value is between zero and one.

7. The method of claim 1, wherein the input data set comprises transaction data.

8. The method of claim 1, wherein the input data set comprises event data.

9. A computer-based method of mining one or more patterns in an input data set of items, the method comprising the steps of:

obtaining an input data set of items;

searching the input data set of items to identify one or more sets of items in the input data set as one or more patterns based on respective comparisons of conditional probability values associated with each of the one or more sets of items to a predetermined threshold value; and

outputting the one or more identified patterns based on results of the comparisons.

10. The method of claim 9, further comprising, prior to the searching step, the step of normalizing the input data set.

11. The method of claim 10, wherein the input data set comprises event data and the normalizing step comprises transforming at least a portion of the event data into event classes such that the event data is non-application-dependent.

12. The method of claim 11, wherein the event data transformation step further comprises the step of mapping two or more attributes associated with an event into an event class.

13. The method of claim 12, wherein the mapping step is performed in accordance with a lookup table.

14. The method of claim 11, wherein the event data is in a tabular form with a first number of columns before the transformation step and in a tabular form with a second number of columns after the transformation step, the second number of columns being less than the first number of columns.

15. The method of claim 9, wherein the outputting step further comprises converting the one or more identified patterns into a human readable format.

16. The method of claim 9, wherein the searching step further comprises the step of performing a level-wise scan based on a set length to determine candidate sets of items in the input data set that have conditional probability values respectively computed therefor that are not less than the predetermined threshold value.

17. The method of claim 16, further comprising the step of pruning candidate sets based on an upper bound property.

18. The method of claim 17, wherein the upper bound property specifies that only candidate sets are considered where the conditional probability of a set of items minus a particular subset of items given the particular subset of items is not greater than the number of occurrences of the set of items minus the particular subset of items divided by the number of occurrences of the subset of items.

19. Apparatus for mining one or more patterns in an input data set of items, the apparatus comprising:

at least one processor operative to: (i) identify one or more sets of items in the input data set as one or more patterns based on respective comparisons of conditional probability values associated with each of the one or more sets of items to a predetermined threshold value; and (ii) output the one or more identified patterns based on results of the comparisons; and

a memory, coupled to the at least one processor, which stores at least one of the input data set and the one or more identified patterns.

20. The apparatus of claim 19, wherein the identifying operation further comprises identifying a set of items in the input data set, which includes at least two subsets of at least one item, as a pattern when the set of items has a conditional probability value computed therefor that is not less than a predetermined threshold value, wherein the conditional probability value is indicative of a probability that both of the at least two subsets of at least one item will occur given that one of the at least two subsets of at least one item has occurred.

21. The apparatus of claim 20, wherein the predetermined threshold value is between zero and one.

22. The apparatus of claim 20, wherein the conditional probability value is estimated by the number of occurrences in the input data set of the two subsets divided by the number of occurrences in the data set of one of the two subsets.

23. The apparatus of claim 19, wherein the identifying operation further comprises identifying a set of items in the input data set as a pattern when the set of items

has a conditional probability value computed for the set of items minus a particular item of the set, given the particular item of the set, that is not less than a predetermined threshold value.

24. The apparatus of claim 23, wherein the predetermined threshold value is between zero and one.

25. The apparatus of claim 19, wherein the input data set comprises transaction data.

26. The apparatus of claim 19, wherein the input data set comprises event data.

27. Apparatus for mining one or more patterns in an input data set of items, the apparatus comprising:

at least one processor operative to: (i) obtain an input data set of items; (ii) search the input data set of items to identify one or more sets of items in the input data set as one or more patterns based on respective comparisons of conditional probability values associated with each of the one or more sets of items to a predetermined threshold value; and (iii) output the one or more identified patterns based on results of the comparisons; and

a memory, coupled to the at least one processor, which stores at least one of the input data set and the one or more identified patterns.

28. The apparatus of claim 27, wherein the at least one processor is further operative to, prior to the searching operation, normalize the input data set.

29. The apparatus of claim 28, wherein the input data set comprises event data and the normalizing operation comprises transforming at least a portion of the event data into event classes such that the event data is non-application-dependent.

30. The apparatus of claim 29, wherein the event data transformation operation further comprises mapping two or more attributes associated with an event into an event class.

31. The apparatus of claim 30, wherein the mapping operation is performed in accordance with a lookup table.

32. The apparatus of claim 29, wherein the event data is in a tabular form with a first number of columns before the transformation operation and in a tabular form with a second number of columns after the transformation operation, the second number of columns being less than the first number of columns.

33. The apparatus of claim 27, wherein the outputting operation further comprises converting the one or more identified patterns into a human readable format.

34. The apparatus of claim 27, wherein the searching operation further comprises performing a level-wise scan based on a set length to determine candidate sets of items in the input data set that have conditional probability values respectively computed therefor that are not less than the predetermined threshold value.

35. The apparatus of claim 34, wherein the at least one processor is further operative to prune candidate sets based on an upper bound property.

36. The apparatus of claim 35, wherein the upper bound property specifies that only candidate sets are considered where the conditional probability of a set of items minus a particular subset of items given the particular subset of items is not greater than the number of occurrences of the set of items minus the particular subset of items divided by the number of occurrences of the subset of items.

37. An article of manufacture for mining one or more patterns in an input data set of items, the article comprising a machine readable medium containing one or more programs which when executed implement the steps of:

identifying one or more sets of items in the input data set as one or more patterns based on respective comparisons of conditional probability values associated with each of the one or more sets of items to a predetermined threshold value; and

outputting the one or more identified patterns based on results of the comparisons.

38. An article of manufacture for mining one or more patterns in an input data set of items, the article comprising a machine readable medium containing one or more programs which when executed implement the steps of:

obtaining an input data set of items;

searching the input data set of items to identify one or more sets of items in the input data set as one or more patterns based on respective comparisons of conditional probability values associated with each of the one or more sets of items to a predetermined threshold value; and

outputting the one or more identified patterns based on results of the comparisons.